Predicting Industry Output with Statistical Learning Methods

November 2017

Peter B. Meyer[1], Wendy Martinez[2]
[1] Office of Productivity and Technology, U.S. Bureau of Labor Statistics
[2] Office of Survey Methods Research, U.S. Bureau of Labor Statistics

Abstract
The U.S. Bureau of Labor Statistics uses estimates of industry output to construct preliminary annual productivity statistics for U.S. manufacturing industries. The official measures of output become available much later. We examine how well several alternative models predict output for each of the 21 industries which make up U.S. manufacturing, using data sources available within four months of the end of the reference year. To measure prediction quality, we use a form of year-wise cross-validation in which industry output from any of the years 2007-2014, even the first, can be predicted by the others and calculate the implied prediction error. This error metric enables us to test the prediction method on each industry over eight years. Our predictors are highly correlated, which reduces the accuracy of prediction. Several methods to address the collinearity problem generate more accurate out-of-sample estimates than ordinary least squares regression. We find that selecting only the best three regressors reduces prediction error by 15%, and a principal components regression by about 20%, compared to OLS.

Key Words: Prediction; industry output; collinearity; cross-validation

1. Introduction

The Bureau of Labor Statistics (BLS) Office of Productivity and Technology produces annual measures of labor productivity for U.S. manufacturing industries. Labor productivity shows the growth rate rates of output per hour worked in the industry. Industry output measures are drawn mainly from sources which are not available until 11 months after the reference year has ended.

This study examines the quality of preliminary estimates of output for each detailed manufacturing industry prepared from a regression analysis of data that is available within four months of the end of the year. There is a tradeoff between timeliness and accuracy, as more data becomes available over time.

The next sections describe the definitions, data sources, and the linear regression methodology which predicts a value-of-production from predictors. The predictors are highly collinear, which produces some overfitting of the regression, and therefore larger errors in prediction. We experiment with variants of linear regression designed to adapt to collinearity: ridge, lasso, elastic net, subset selection, and principal component regressions. The best-performing of these methods in the out-of-sample tests here are subset selection to three predictors which improved accuracy versus ordinary least squares regression by 15%, and principal-components regression which improved on OLS by 20%.

2. Definition of industries, output, and growth rates

U.S. manufacturing establishments are classified by their primary products into categories of the North American Industry Classification System (NAICS). There are 21 3-digit NAICS classifications in manufacturing, with numbers ranging from 311 to 339. Each of these 3-digit industries is subdivided into 4-digit industry categories, with numbers ranging from 3111 to 3399. All of our data sources use these classifications, but sometimes the published statistics group together some of the 4-digit industries. The main results below are for 3-digit industries. Each industry-year is an observation in the regressions to follow.

The final data for industry output comes from the Census Bureau's Annual Survey of Manufactures (ASM) and are not available until approximately 11 months following the end of the reference year. In years ending in 2 or 7, the more complete Economic Census is conducted, and the ASM is not conducted. For these years, the final industry output figures are not available for more than 18 months after the reference year. The 2012 Economic Census became available in late 2014.

The output measure to be predicted for each industry is its annual **Value of Production** in current (nominal) dollars. Value of Production is a sectoral output measure, meaning it includes production that is sold or transferred outside the industry but excludes production that is sold within the industry. Industries are made up of establishments, and if an establishment produces goods in multiple industry categories, all its revenue is counted in the largest one, its primary industry.[1]

The Value of Production is constructed from four components, each of which is expressed in current dollars: the sum of *shipments* and *change in inventory*, minus *resales* (goods resold without transforming them) and *intrasectoral transactions* (sales within the industry). The first three elements are drawn from the Economic Census or the Annual Survey of Manufactures. The fourth component, intrasectoral transactions, is estimated by BLS/OPT based on Economic Census data describing the flow of products between establishments within the same industry.[2]

Annual growth rates of each industry's value of production (VoP) are defined thus:

$$\text{Growth rate of VoP in year t} = \ln\left(\frac{VoP_t}{VoP_{t-1}}\right) \qquad (1)$$

[1] Revenue from an establishment's products that fit into its industry category is called *primary output*; revenue from products fitting more naturally into other industries is called *secondary production*. The basic chemicals industry (NAICS 3251) has the highest fraction of secondary production – about 14 percent of its output. The presence of secondary production creates a minor mismatch between some of the predictor variables and the output measure, since some of these variables are associated with establishment activity and others with the primary industry. The mismatch is not explored further in this paper, but it would be possible to construct predictions separated along this dimension.

[2] *Intrasectoral transactions* are flows of goods and payments between establishments within an industry (Gullickson, 1995, and https://www.bls.gov/news.release/history/prod2_08082006.txt).

This unitless logarithmic growth rate is a standard construct in such regressions. In this work, growth rates of the dependent variable and the predictors were all calculated in this form.[3]

3. Empirical summary of the output growth rates

The overall growth of value of production across all 21 manufacturing industries over this period was 1.2 percent per year. The unweighted average growth, however, is near zero because growth was concentrated in certain large industries whereas many of the small industries were shrinking. In several cases, output rose or fell by 50 percent from the year before, notably in some of the primary metals production industries (NAICS 331) in the turbulent year 2009.

The data needed to construct the value of production for each industry are available from the ASM, which is usually released in November following the reference year. These values are typically revised in the following year's ASM.

In 2011, 3-digit industries varied from $5.665 billion in annual revenue (for NAICS 316, Leather and allied products) to $838 billion (for NAICS 324, Petroleum and coal products). Four-digit industries varied from $1.44 billion in annual revenue (for NAICS 3159, Apparel accessories) to $838 billion (for NAICS 3241, Petroleum and coal products). Predictions tend to be more accurate for larger industries, presumably because the predictors and the estimates of final output are both constructed from larger samples.

4. Data Available to make Preliminary Estimates

Several data series associated with industry activity levels are available early enough so that we can use them to predict the value of production for the previous year. Their year-to-year changes are positively correlated to output changes. The subsections below discuss each predictor variable, the industries it covers, and the timing of its availability.

4.1 Industrial Production Indexes (IPIs)

The Federal Reserve Board releases estimates of real output by industry each month. These are called the Industrial Production Indexes. For 50% to 60% of manufacturing industries, the IPI draws mainly from data on physical quantity output from trade associations and government agencies and from estimates of industry activity from the Census Bureau's Capacity Utilization Survey. When such physical quantity measures are not available, IPI estimates are based mainly on production hours worked in the industry, which are drawn from BLS's Current Employment Statistics (CES). IPI estimates draw from industry-specific regressions and judgment by industry specialists at the Federal Reserve.[4]

The IPIs cover all 3-digit and 4-digit NAICS manufacturing industries. However, NAICS industries 3311 (iron and steel mills and ferroalloy manufacturing) and NAICS 3312 (steel

[3] Experiments with the data in arithmetic growth rates of the form $(X_t-X_{t-1})/X_{t-1}$ did not produce better predictions.

[4] Sources: Communications with Kim Bayard and Charlie Gilbert of the FRB, from http://www.federalreserve.gov/releases/g17/About.htm and from the Industrial Production Explanatory Notes, Industrial Production and Capacity Utilization - G.17, at the Federal Reserve Board web site. http://www.federalreserve.gov/releases/g17/IpNotes.htm.

product manufacturing from purchased steel) are combined into one index. Indexes for some 4-digit industries are available only on request.

The first IPI report for a given month is released about six weeks after the end of that month and is revised in later releases. The first annual estimates for a reference year are available by the end of January and are updated in succeeding months.

IPI growth rates are later benchmarked to the Economic Census and the ASM. This study uses IPI growth rates for each year that were released before this "benchmarking" was conducted, since in the context of preliminary estimates, ASM and Economic Census data are not yet available for the reference year.

4.2 Producer Price Indexes (PPIs)

Manufacturers are producers, and *producer price indexes* deflate the value of the items they produce, whether sold to other producers or to consumers. PPIs are helpful for this model because some variables are nominal, others real. The dependent variable, value of production, is in current (nominal) dollars. Several predictors, such as the IPI indexes, are in constant-dollar terms. There are PPIs for every 3- and 4-digit industry, matched with products as accurately as possible.

Monthly PPIs are published three weeks after the end of the reference month, and annual averages are available at the end of January. Small revisions based on corrections and late respondents are published four months later. By the end of March, near-final PPIs are available for the reference year.

4.3 M3 shipments data

The Census Bureau's Manufacturers' Shipments, Inventories, and Orders (M3) data are based on a voluntary monthly survey of U.S. manufacturing companies, including most companies with $500 million or more in annual shipments. M3 data include the value of shipments, total inventories, inventories by stage of fabrication, new orders received, and unfilled orders.

The M3's accounting concepts are the same as those of the ASM, but the M3 survey varies from the ASM in other respects: the M3 covers companies, not establishments; it does not cover all 4-digit manufacturing industries; the sample size is smaller than in the ASM, and the M3 survey has a low response rate. Large companies are overrepresented in the M3 sample by selection, and the Census Bureau reweights the responses to represent the size distribution of the population of actual companies.

The M3 survey covers all 3-digit NAICS and roughly 23 of the 4-digit NAICS manufacturing industries. The 4-digit coverage includes some combinations and partial coverage, which we stretch into proxy measures by subtracting and dividing the known numbers. Estimates for the remaining 4-digit industries are constructed by subtracting the known 4-digit estimates or proxies from the 3-digit industry that contains them, as explained in the section "Constructing proxies." The term "proxy" is used for the M3 values because they do not perfectly match the industry.

The M3 survey is available shortly after the end of each month. Preliminary and revised year-end results of the M3 survey are typically released two and three months after the end

of the calendar year, so the revised data are available in time to be used to generate preliminary estimates.

4.4 Imports and exports data

Monthly data on imports and exports by industry is gathered by the U.S. Customs and Border Protection (CBP) and made available shortly after the end of the month by the U.S. International Trade Commission (USITC). Imports are categorized by the kind of product that was imported and the products can be mapped to industry categories.[5] In this way, the USITC imports and exports data covers all 3- and 4-digit manufacturing industries except one, for which a good proxy is available.[6]

4.5 Wages and employment data

The Quarterly Census of Employment and Wages (QCEW) covers over 96 percent of manufacturing-industry jobs in the United States. The QCEW data includes employment counts and the total wage bill for each 4-digit manufacturing industry, and we use these as separate predictors here.

QCEW data is available about five and a half months after the end of each quarter. Near the end of March, the data for the third quarter of the previous year becomes available. We therefore use the growth in wages and in employment from the first three quarters of one year to the first three quarters of the next year. It is feasible to use other data to construct estimates for the fourth quarter of the reference year, but our past experiments using such estimates as predictors did not improve predictions.

4.6 Construction of "proxy" predictors for four-digit industries

In some data sources, such as the M3, certain industries have been combined, or are not included. If two industries are combined in a statistical release, we impute the growth rate of the total to the component industries. This kind of estimate has worked well as a proxy. For most 4-digit industries which are not reported, proxies are computed by taking the best available 3-digit total and subtracting the component 4-digits that are observed. The remainder approximates the best available growth rate for this 4-digit element.

---

[5] Imported and exported products are classified by "end-use code," associated with categories such as corn, medicinal equipment, or civilian aircraft. These categories are documented on the Census Bureau's Foreign Trade Reference Codes web page: http://www.census.gov/foreign-trade/reference/codes/index.html#enduse. Goods are also classified into the Harmonized System Codes (HS Code), discussed at http://www.foreign-trade.com/reference/hscode.htm. BEA further estimates the proportions of imports which are put to use as materials in production, versus long-lasting investment capital goods, versus final consumption, and these differences are used in productivity statistics. In this study, we use the values of imports, not information on how they were used.

[6] The missing one is NAICS 3328 (Coating, Engraving, Heat Treating, and Allied Activities). For that industry, the annual changes from imports and exports from NAICS 3327 (Machine shops; turned product; and screw, nut, and bolt manufacturing) worked well to predict output growth in 3328. Other sub-industries and residuals from industry 332 were considered, but 3327 performed best. After this substitution, estimates of imports and exports levels at the 4-digit level no longer add to the estimated 3-digit levels, but only growth rates, not levels, matter in the regressions.

The same issue arises for other data sources, to a lesser degree. Import/export data is missing for industry 3328, but growth rates of industry 3327 were found to be a good proxy to predict output growth in 3328. The IPI combines 3311 and 3312 data, and the movements in the total were good proxies for each. In the future, we may use sales and revenue figures from the Census Bureau's Quarterly Financial Reports (QFR), which are good predictors for industry output. Because the QFR data do not cover all the industries, and its classifications have changed over the years (Census Bureau, 2012), the QFR is not used in this study.

4.7 Summary of predictors and regressions

The regressions below use the growth rates of all seven measures of industry activity discussed above as independent variables. Including industry indicators and past output levels as regressors did not consistently improve out-of-sample performance, and they are left out of the regressions shown here.

Within three months of the end of the reference year, seven of these predictors are available, at least in preliminary versions. An additional month would not give much more data. These tests therefore use data as it was available before each April when possible, and do not include later revisions of the predictors.

Figure 1 illustrates the scale of the growth rates, shown in arithmetic form (X_t / X_{t-1}). Each of the four main columns is numbered for a NAICS 4-digit industry. Data points in the left column of the pair of columns above each industry show growth rates of the predictor variables, and the data points on the right show the dependent variable and one model's prediction of them. Generally, the predictions here are averages of the predictors. For industry 3112, the variable to be predicted is outside the range of the predictors. If this is a consistent pattern, a model could accommodate it with a constant, a fixed effect by industry or other predictors.

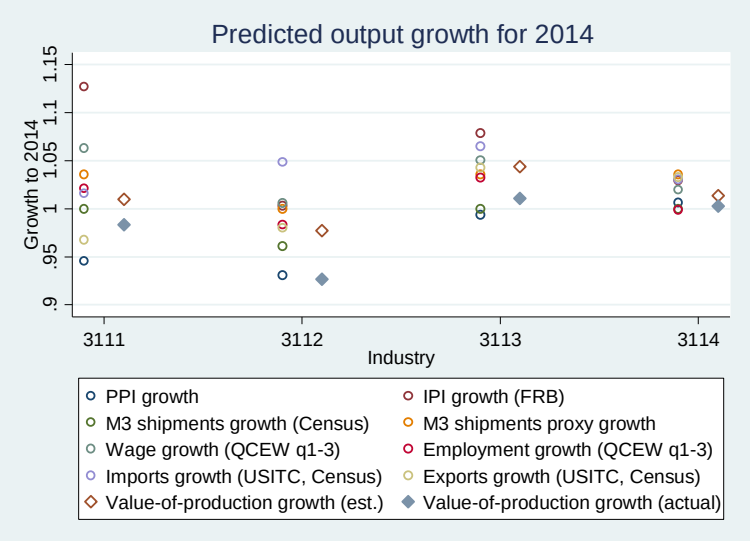

Figure 1: Growth rates of example predictors, industry output, and predictions of output

4.8 Quality of these predictors

Growth in each of the predictors is highly correlated with the main dependent variable, the growth in value of production, at both the 3-digit and 4-digit level. Correlations are greater at the 3-digit level than the 4-digit level, presumably because samples are larger in larger industries.

Table 1 summarizes the results of single-variable regressions used to estimate the growth rate of the value of production. Each line is from a separate regression. The specification of these log-linear regressions for industry *i*, year *t*, and predictor X is

$$\ln\left(\frac{vp_{i,t}}{vp_{i,t-1}}\right) = \alpha + \beta \ln\left(\frac{X_{i,t}}{X_{i,t-1}}\right) + \varepsilon \qquad (2)$$

Table 1. Univariate regressions of industry value-of-production growth, 2007-2014
N=168 industry-years

Predictor	**Regression coefficient**	**R^2**
IPI	1.19	0.59
PPI	1.31	0.36
M3 shipments	1.08	0.71
Imports	0.76	0.76
Exports	0.78	0.71
Wages, quarters 1-3	1.32	0.59
Employment, quarters 1-3	1.51	0.49

5. The collinearity problem

The predictors are highly collinear. A core problem here is that collinearity of the predictors leads to overfitting of the regressions, which creates predictions that are overly sensitive to random variation in the predictors. Pair-wise correlations across the predictors are high, as shown in Table 2.

Table 2. Pair-wise correlations across the growth predictors

	IPI	PPI	Imports	Exports	Wages	M3
IPI	1.000					
PPI	0.207	1.000				
Imports	0.780	0.600	1.000			
Exports	0.599	0.648	0.824	1.000		
Wages	0.853	0.283	0.728	0.638	1.000	
M3 shipments	0.731	0.647	0.865	0.776	0.736	1.000
Employment	0.793	0.198	0.614	0.545	0.946	0.647

Pairwise correlations do not perfectly indicate how much multicollinearity will affect a regression. The **Variance Inflation Factor** (VIF) measures how much the variance of each OLS-estimated coefficient was raised by collinearity. The VIF for each predictor *i* is computed from the R^2 from regressing it on the other predictors:

$$VIF_i = \frac{1}{1-R_i^2} \quad (3)$$

According to standard heuristics, a regression VIF over 4 calls for attention, and a VIF over 10 calls for particular scrutiny. Table 3 shows VIFs in our data.

Table 3. Variance Inflation Factors

	3-digit industries	4-digit industries
IPI	6.41	2.79
PPI	3.01	1.67
Imports	7.64	3.05
Exports	3.74	2.12
Wages	15.95	8.01
Employment	11.07	7.12
M3 shipments	5.74	2.44

Our goal is to build a model that is not too sensitive to collinearity, avoids overfitting the data, and produces good prediction accuracy. Pruning variables would be feasible, but our objective in this paper is to evaluate methods that make some use of the unique information in all these variables.

6. Measures of prediction accuracy

Measures of the accuracy of a prediction method are somewhat different from those that measure the fit of a regression. A regression fits well to the degree that the independent variables jointly correlate to the dependent variable within the sample. However, the practical problem to be addressed is to use past years of full data to predict the output in the target year for which output data is not available. In the language of data science, the prediction problem calls for the use of a "training data set," for which the right answer is known, to calibrate the exact prediction model, including the coefficients of any regression. The quality of the method is then judged on the accuracy of its predictions on "test data" which was not included in the training data but is believed to have come from the same data-generating process, and to have the same relationships between the variables. Measures of accuracy in which the training data and test data are separate will be called *out-of-sample* measures of accuracy.

In practice, we will have near-complete data for all industries on a set of past years and would want to predict all industries for the latest year. Simulating that problem in these tests, we could repeatedly attempt to predict the latest year with full data, 2014, and compare accuracy across methods. However, 2014 data may be distinctive, and we can obtain more information on the accuracy of methods by treating each year as test data. Therefore, for each method we carry out the following form of cross-validation.

We treat observations from 2007-2013 as the training data, run the regression or other prediction method to "train" the coefficients, and apply the resulting model to 2014 data for each industry. For each 2014 industry, we measure the out-of-sample error of this method by the difference between the predicted value and the actual industry output observed later. We denote this difference, this error or residual, by e_{it}, where i is industry and t is year. Then, we conduct this exercise again using 2007-2012 and 2014 together as the training set and 2013 as the test set, and so on, using each year as test data once. Each year's data will be test data in one "fold," in this k-fold construction.

At the conclusion of this process, for each method we have a set of residuals e_{it} for each industry and year, and we can use them to evaluate the performance of the method overall against alternative methods, and by year or by industry. The term "cross-validation" describes the construction of out-of-sample errors of this kind. Because we did not drop one observation at a time (the usual kind of cross-validation), but a whole year of data at a time, this is "year-wise cross-validation." It would be possible to drop one observation at a time and compute coefficients from the others, including those from the same year, but this would not reflect as well the problem to be faced in production; the error rates would be biased low because more information would be available in the test than in the production environment.

A panel of errors e_{it} for each model can be described by several different statistics to characterize the performance of the model compactly:

- R^2, the coefficient of determination, which is a unitless number between zero and one expressing the fraction of the total sum of squared deviations in the dependent variable that is fit by the regression or other prediction method. An out-of-sample version of this statistic can be computed, but in the end we will not show R^2s here because it does not translate quickly into a measure of the level of error the user of the final statistic, the dependent variable, will experience. To envision this, imagine a panel of industry-years where the predictors have almost no relation to the output levels, but output varies only slightly each year. The regression then performs poorly but the predictions could be good because the total deviations from year to year are small. In extreme cases, an R^2 could be near zero or undefined even when predictions are adequate.

- The maximum observed error for each model across all industries, all years, or all industry-years. In general, the more economically turbulent earlier years 2007-2010 have larger prediction errors than the later years, and small industries have larger prediction errors than large ones, perhaps because of sample sizes in the predictors.

- The mean absolute error (MAE) is the absolute value of the difference between the predicted growth rate and the actual growth rate: $MAE = \frac{1}{n}\sum_{i=1}^{n} |e_{it}|$ This measure has the same units as the dependent variable. Here, an MAE of .02 represents an average of 2% error in the dependent variable, which is an industry output growth rate. The MAE and RMSE have the advantage that they are always defined for any set of predictions and will therefore always give a ranking of prediction methods.
- Root mean square error: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} {e_{it}}^2}$ This metric matches the optimality criteria of ordinary least squares regression. RMSE, unlike MAE, penalizes errors that have a large absolute value more than a set of smaller ones with the same total. RMSEs

are in the same units as the dependent variable. RMSEs are larger than MAEs, on the order of 40% in this data, and grow larger yet with larger sample sizes.

We have chosen to show the RMSE and not the other statistics in this paper for simplicity and to match the usual methods in the literature. Chai and Draxler (2014) discuss the choice between MAE and RMSE. They report that the MAE is more suited to uniformly distributed errors, whereas the RMSE is more appropriate if the errors have a normal distribution, as they seem to here. In the data and models shown here, the correlation between the RMSE and the MAE is approximately .98. We show the RMSE in comparisons across methods below.

We have checked for time series relations among the growth rates, and we do not find trends over time in how they relate to one another, which would make prediction of earlier years in this way more dubious.

7. Models tested

We tried several modeling approaches. These include ordinary least squares (OLS), ridge regression, Lasso, elastic net, subset selection, and principal component regression. See Hastie, Tibshirani, and Friedman (2009) for an excellent introduction to these methods.

We used the R computing environment (https://cran.r-project.org/) and the `caret` package (http://topepo.github.io/caret/index.html) for this analysis. The `caret` package provides an interface to many modeling functions in R, making our analysis consistent across the various approaches.

7.1 Ordinary least squares

Ordinary least squares estimates minimize squared-errors and are unbiased:

$$\hat{\beta}_{OLS} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{N} (y - X\beta)^2 \right\} \quad (4)$$

When these coefficient estimates were constructed in the 3-digit industry sample sets and applied to make predictions in the out-of-sample year, the average RMSE was .082, that is, 8.2% error in output growth. Residuals from OLS appear normally distributed and appear to be autocorrelated. We do not have an explanation of why they are autocorrelated.

Coefficients vary somewhat in the yearly models obtained through cross-validation. Among the systematic relationships are those of employment and wages. Employment change and wage change are highly correlated to one another. When both are included in regressions predicting output, employment has a positive coefficient and wages have a negative coefficient. This makes sense insofar as employment increases directly cause more output, whereas wage increases make creation of output more costly. Relatedly, expanding industries tend to hire low-wage staff at the margin, whereas industries which are contracting in size tend to keep relatively high-wage employees and not to hire new low-wage staff. Thus employment counts are correlated more highly to output than wages are.

Table 4. OLS predictors of VoP growth all 3-digit manufacturing industries
All variables are in log growth form: ln(value in year t / value in year t-1).
N=168 (21 3-digit industries for 2007-2014. R^2 = 0.69, RMSE=.073 (in sample)

	Coefficient	p-value
PPI growth	0.295	0.086
IPI growth	0.307	0.083
M3 shipments growth	0.029	0.833
Imports growth	0.211	0.052
Exports growth	0.314	0.000
Wage growth, quarters 1-3	-0.704	0.006
Employment growth, quarters 1-3	1.11	0.000

7.2 Ridge, Lasso, and Elastic Net regressions

A **ridge regression** is a constrained least-squares regression. This method is designed to reduce sensitivity resulting from collinearity by pressing regression coefficients toward zero. The ridge regression equation can be expressed in different forms, with equivalent constraints. For *p* parameters and *N* observations and an R chosen by the investigator:

$$\hat{\beta}_{ridge} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{N} (y - X\beta)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq R \quad (5)$$

One form of the ridge regression constraint is denoted R above. R is a hyperparameter generally selected by cross-validation from the training data. We let the constraint be chosen automatically by the `caret` package's cross-validation and training functions. The resulting RMSE for the 3-digit industry panel was 0.077. Thus ridge regression slightly outperforms OLS.

A **lasso regression** is an alternative constrained least-squares regression method. Coefficients are constrained by the sum of their magnitudes.

$$\hat{\beta}_{lasso} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{N} (y - X\beta)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq L \quad (6)$$

The lasso method tends to select among regressors by pushing some of the estimated coefficients toward zero (Tibshirani, 1996). In this application, the results are about the same as for ridge regression, with an RMSE of 0.077.

The **Elastic net** method incorporates both ridge and lasso constraints, weighing the R and L constraints above by performance in cross-validation. It performed better than the earlier methods, giving an RMSE of 0.074.

7.3 Best three predictors – subset selection

Three predictor variables tended to have the highest importance. Furthermore, they are not very highly correlated with one another. These are IPI, PPI, and Exports. Using only these three predictors in an OLS regression, we obtained an out-of-sample RMSE of 0.070.

7.4 Principal-components regression

In principal-components regression, one first applies principal component analysis to the predictors and then uses these constructed variables, the components, in OLS to make predictions. We obtain uncorrelated observations when we transform our data to the space spanned by the principal components, which is one way to address collinearity. We could keep all principal components or use a smaller number, which reduces the dimensionality of the regression. Typically, one would retain enough components to explain a high percentage of the variation in the original data (Jackson, 1991).

The first principal component is a linear combination of the 7 growth predictors, and we found that only one principal component was needed to explain over 95% of the variation. The first component weighted most heavily the IPI, PPI, exports, and employment-minus-wages. Estimates were produced by the `pls` package in R (Mevik, Wehrens, and Liland, 2016).

The first component was then used to predict VoP growth by OLS. The RMSE for 3-digit industries with this method, using one component, was 0.065 -- the best of all the methods we tried. To our surprise, adding the second principal component as a regressor did not reduce the prediction error.

8. Discussion, alternatives, and future research

We have experimented with splines and random forest models, but these methods did not improve significantly on the performance of OLS. We think the reason is that the underlying problem doesn't call for their flexibility. The modeled relationships seem basically linear, with no major bends and curves. The flexibility enabled by these methods did not help accuracy in prediction for this problem. One reason could be that the ratio of numbers of observations to the numbers of parameters, sometimes expressed as N/P, is not large enough. Here there are N = 147 observations in training set for 21 3-digit industries. Furthermore, dividing up industries into groups to create industry indicators did not seem to help accuracy. Estimates for some more complicated models were not computable because there were too few observations.

Table 5 summarizes the performance of the available model predictions. Random forests predictions vary slightly from run to run, and the figure shown is an average.

Table 5. Summary of model performance metrics

Model	Root mean squared error (RMSE)
Ordinary least squares with all 7 variables	0.082
Random forests	0.080
Ridge regression	0.077
Lasso regression	0.077
Elastic net	0.074
OLS with best 3 predictors	0.070
Principal components regression	0.065

Compared to OLS with all seven growth variables, we find in these tests that pruning variables to the selected three reduces out-of-sample prediction errors by approximately 15%, and principal components regression using all seven reduces errors by approximately 20%. For methods other than OLS, generally speaking, the in-sample performance of the regression is slightly worse if a regressor is eliminated.

The use of wage and employment predictors in forecasting output generally produces slightly better output estimates, mainly by reducing the most extreme errors. If the resulting estimate is used to produce a productivity measure, there is a beneficial side effect to using these predictors. When errors in labor estimates are large and positive, they also raise the predicted output level, biasing it high, and so the effect of such errors is muted in the resulting productivity ratio. When errors in employment and wages are low, they lead to an under-prediction of output. Thus, there is a negative feedback effect of incorporating errors in labor estimates which tend to slightly reduce errors in early productivity estimates.

The fact that the principal components method works best suggests that the underlying the data-generating process has a particular pattern. These measures of industry activity – value of production, shipments, employment, wages, and the like – are related entirely linearly, and perhaps they are close to being measures of the same underlying flow, and there are no curvilinear or threshold relations among them.

There are a number of additional experiments which may improve predictions further:

- There is autocorrelation in the residuals from the OLS regression, and using these residuals as predictors could therefore improve predictions. It is not clear why there is autocorrelation.
- Interactions between the predictors might create better predictors.
- It is possible to use these and other predictors to estimate each of the four components of the Value of Production -- shipments, inventory change, resales, and intrasectoral transactions. The accounting rules require certain components of four-digit industries to add exactly to their levels in the three-digit industries, and this fact can be applied to slightly improve predictions, as has been shown in past research (Meyer et al. 2015). To do this here would add complexity to our evaluation of the more advanced predictions methods, so in this study we show only direct predictions of the Value of Production.
- Evidence from across industries is sometimes best modeled with a random coefficients model, a linear model in which coefficients can have a normal distribution, varying somewhat by industry.
- Finally, it would be possible to use other data sources in the regression. Data from the Census Bureau's Quarterly Financial Reports is available for many of these industries, treating reported revenues or profits as predictors of output. The American Association of Railroads releases estimates of transport volumes and values for the products of U.S. manufacturing industries. These data sources are organized by NAICS industry and are available early enough to be included in the predictions here, but data limitations and required transformations made them too difficult to apply in this work.

9. Conclusion

The regression approach can include many variables for predicting value-of-production levels by industry. For large industries, it is possible to get good estimates four months

after the reference year ends, but not for all industries. Prediction accuracy is verifiably affected by the high collinearity of the predictors. This collinearity leads to regression coefficients that seem to be excessively sensitive to small variations in the predictors, according to our out of sample tests.

We explored several regression methods to reduce the problem, and measured their performance by year-wise cross-validation. We find that compared to OLS with all seven growth variables, selecting the best three predictors reduced out-of-sample prediction errors by 15%. Principal components regression reduced errors by 20%.

Acknowledgements

This work is drawn from work on a long-term project to improve preliminary output and productivity measures conducted in cooperation with BLS colleagues, including Jennifer Kim, Jason Dale, Jason McClellan, Jennifer Price, Sam Rowe, Jim Mildenberger, Michael Brill, Bobbie Joyeux, and Mike Manley. The authors thank Leo Sveikauskas, Jay Stewart, Yvonne Taylor, and Kelly McConnville for valuable advice. Views expressed are those of the authors not the Bureau of Labor Statistics.

References

BLS. PPI release dates. http://www.bls.gov/schedule/news_release/2014_sched.htm

Census Bureau. March 2012. *Quarterly Financial Report for Manufacturing, Mining, Trade, and Selected Service Industries*, 2011 Quarter 4. http://www2.census.gov/econ/qfr/pubs/qfr11q4.pdf

Census Bureau. March 26, 2014. EC1200CADV1: All sectors: Core Business Statistics Series: Advance Summary Statistics for the U.S. (2012 NAICS Basis): 20122012 Economic Census of the United States**.** http://www.census.gov/newsroom/releases/pdf/2012_econ_advance_report.pdf.

Chai, T.; R.R. Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, *Geoscientific Model Development* 7:3, 1247-1250, doi:10.5194/gmd-7-1247-2014

Federal Reserve Board. Industrial Production and Capacity Utilization measures. http://www.federalreserve.gov/releases/g17/About.htm and http://www.federalreserve.gov/releases/g17/IpNotes.htm

Gullickson, W. 1995. Measurement of productivity in U.S. manufacturing. *Monthly Labor Review*. https://www.bls.gov/mfp/mprgul95.pdf

Hastie, T.; R. Tibshirani; J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag. https://web.stanford.edu/~hastie/ElemStatLearn/

Jackson, J. E. 1991. *A User's Guide to Principal Components*, John Wiley and Sons.

Mevik, B-H.; R. Wehrens; K. Hovde Liland. 2016. Pls package for R, version 2.6-0.

Meyer, P. B.; J. McClellan; J. Price; S. Rowe; J. Mildenberger; M. Brill; J. Kim. Early estimates of annual manufacturing industry output. Presentation at the Federal Committee on Statistical Methodology conference, 3 Dec. 2015.

Stock, J.; M. Watson. 2012. Generalized Shrinkage Methods for Forecasting Using Many Predictors. *Journal of Business & Economic Statistics* 30(4): 481-493.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B.* 58:1, 267-288.